Carleton
UNIVERSITY

Canada's Capital University

ARTIFICIAL INTELLIGENCE FOR MOLECULAR DESIGN

by

Benjamin Kachkowski Weiser

A thesis submitted to the Department of Chemistry acceptance in partial fulfillment of the requirements for the degree of

Bachelor of Science with Honours
in
Chemistry and Physics

Carleton Univeristy
Ottawa Ontario

© 2020
Benjamin Kachkowski Weiser

The undersigned hereby recommends to the Department of Chemistry acceptance of this report, submitted by Benjamin Kachkowksi Weiser in partial fulfillment of the requirements for the degree of Bachelor of Science with Honours

Thesis Supervisor

Abstract

The field of chemistry is uniquely positioned to extend molecular technologies to far more advanced levels than current synthetic efforts. Our inability to make large advancements instead of incremental improvement is not a result of a lack of effort. It is due to the limitations of the human mind. We are thus unable to design molecular structures containing hundreds of thousands of atoms, each with a specific purpose. The conclusion is human-made molecular designs would need assistance beyond human capabilities to rival nanochemistry which is found in biology designed through evolution. Therefore, another tool must be used which can better account for those limitations and can design molecules by an alternative process to human creativity. Due to advances in computer science and machine learning (ML) technology, various tools are becoming available. Using ML techniques for quantum chemistry simulation is demonstrating capabilities for making simulation significantly less computationally tasking with ample accuracy for chemical applications. ML is being used for applications ranging from calculating stable structural conformations to coarse-grained models within molecular dynamics simulations. Research into applications of ML for molecular property prediction has shown promising results. Properties regarding the electronic structure, reactivity, toxicity, receptor binding capabilities, protein folding structure from its amino acid backbone, are now routinely predicted by ML programs (Cova and Pais 2019; Senior, Evans, and Jumper 2020). This essay investigates combining ML methods for simulation and molecular property prediction with genetic programming, another type of machine learning, which emulates evolution, to create novel designs. Applying this method to chemistry will allow us to mimic the evolutionary process to create new synthetic designs. This tool would allow for an immense progression in synthetic molecular design resulting in impressive advancements in chemical research.

Acknowledgements

Thank you to my supervisor Dr. Maria DeRosa.

Thank you to Dr. Issaka Seidu for the enormous amount of time and commitment which you put into this project as well as encouraging my excitement. It was a pleasure working with you.

Thank you to my mother for letting me discuss my ideas with her and for the aid in editing my terrible spelling. More importantly, thank you for the constant support and motivation to be as amazing as you.

Contents

1 Machine Learning for Chemistry **1**
1.1 Chemical Simulation 2
1.2 Molecular Property Prediction 6

2 Genetic Programming **7**

3 AI for Molecular Design **8**

4 AI for Molecular Design Applications **12**

5 Future Work **13**

List of Figures

1 Diagram of AI terms 1
2 Neural network and deep neural network schematic 2
3 Quantum-chemical simulation process 3
4 Potential energy surface for water molecule 4
5 Deep ML applications for chemical problems 9
6 AI for molecular design flowchart 10
7 Molegear generated design compared to indinavir 11
8 HIV-1 inhibitor molecules designed by Molegear 11

1 Machine Learning for Chemistry

Machine learning is a glorified term to describe an algorithm that finds patterns in large amounts of data through the minimization of errors in a variety of different ways. The ML minimization algorithms are used by Google for search results, Apple for their voice assistant, and Facebook for personalized site feeds. Machine learning has also been applied to academic research in virtually every scientific field.

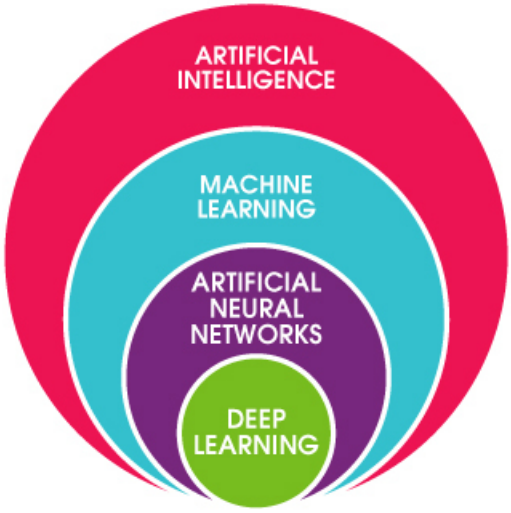

Figure 1: Diagram of AI terms (Hauptfleisch 2016)

As depicted in Figure 1, artificial intelligence (AI) is a broad term which can describe any software which demonstrates capabilities akin to human intelligence. Machine learning is a type of AI that is able to modify its parameters when presented with more data. This means that ML programs are able to self-construct modifications and therefore less reliant on human construction of the program. Neural networks are a type of ML named for the modest similarity to human brain structure and the type of interest to this paper is Deep Learning, see Figure 2. Machine learning algorithms are written to learn the relationship between a data set and some desired property. Therefore, the quality of the relationship obtained depends on the complexity of the neural network(s) employed. The outputs or predictions are products of the inputs and the hidden algorithm of the neural network. Deep learning refers to a neural network that contains multiple hidden layers as shown in Figure 2. The complexity of the hidden layers determines the level of sophistication (mathematical) between the input and output layers. These hidden layers allow for the neural network to process information about the inputs sequentially. Simplifying facial recognition, the first hidden layer can be used to determine lines in a picture, the second layer is used to determine the features which those lines make up identifying eyes, mouth and nose. Lastly, those features are given to the last layer where the identity of the person is determined from those features and an answer is outputted.

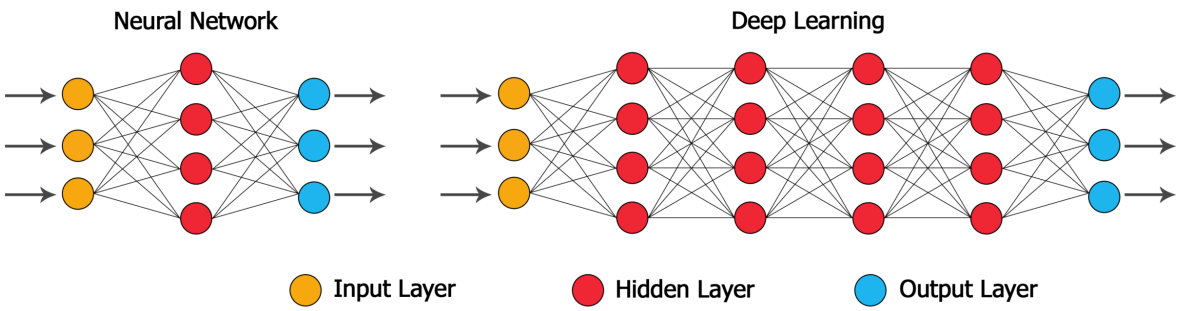

Figure 2: Diagram of neural networks and deep neural networks. The inputs [orange] are sent to the hidden layer [red]. The hidden layer and input yield the output [blue] for the given input. For deep learning, multiple hidden layers are applied to yield the output (alphabold 2019)

The field of machine learning for applications in chemistry is very much in its infancy and the coming years will have working applications that touch upon a wide varying range of chemical problems. ML is being investigated for the following uses, including but not limited to; determining reaction mechanisms, approximations of density functional theory (DFT) functionals, molecular dynamic simulations, reaction outcome prediction, optimal condition prediction, chemical optimization, protein folding structure prediction, improving catalytic technologies, analytic tools, quantum sensing (Rupp et al. 2012; Ahneman et al. 2018; Behler 2016; Cova and Pais 2019; Wang et al. 2019; Shindo and Matsumoto 2019; Senior, Evans, and Jumper 2020; Freeze, Kelly, and Batista 2019; Segler and Waller 2017). There are four ML applications of significant interest to this essay, namely: quantum molecular simulations; molecular property prediction; screening; and development of synthesis procedures for new compounds, and finally, molecular design using genetic programming. The outcome of the complete process is stable molecular designs with desired properties.

1.1 Chemical Simulation

The traditional approach to molecular simulation computational chemistry involves the use of quantum chemistry approaches, in some cases coupled to quantum dynamic techniques. The most widely used quantum chemistry software for this task is DFT because of its compromise between accuracy and computational cost. DFT is an approximate *ab initio* for calculating molecular properties. The result of these calculations is the potential energy surface for a given molecule. These calculations are run over all vibrational degrees of freedom. For this reason, DFT calculations for these kinds of applications are very computationally expensive with today's hardware technology. Run times can be months for larger molecules. Additionally, the accuracy of DFT is affected by the fact that DFT fails in some cases such as accounting for van der Waals forces.

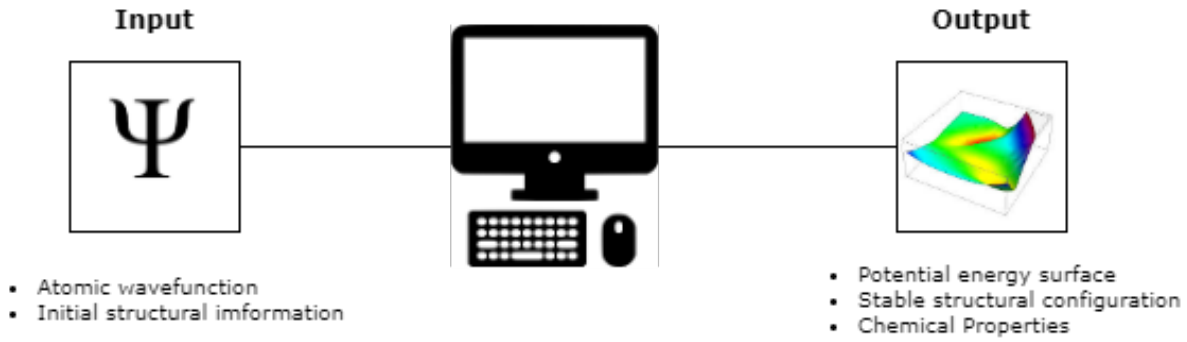

Figure 3: By using initial structural information and individual atomic wave function computer can be used to solve chemical problems. Outputs can yield the PES, stable structure, and thus chemical properties. Due to high computational expense of such simulations, computer clusters are often necessary.

Attempts at substituting quantum mechanical simulations using machine learning have shown significant promise in applications such as emulating DFT calculations, optimizing molecular dynamic simulations, and developing coarse-grain models in molecular dynamics. The expansive set of data already collected by traditional computational methods means there are large data sets available to train ML algorithms for desired applications. The purpose of these simulations is to predict solutions, without explicitly solving Schrodinger's equation, for a given molecular structure resulting in the energy of a molecule at a given structure (nuclear configuration) as well as other properties of the molecule. Calculation of energies along all 3N-6 (3N-5 for linear molecules) vibrational coordinates yields a potential energy surface (PES). Depicted in Figure 4 is the PES of water showing the asymmetric stretch and bending vibrational motions. The resulting multi-dimensional potential energy surface gives structural details such as global and local minimum energy geometries, transition state geometries, and products. It allows for the determination of the relationship between structure and stability, as well as information about transition states. ML algorithm has the capability of learning and predicting complex PES and properties with minimal effort and computation cost.

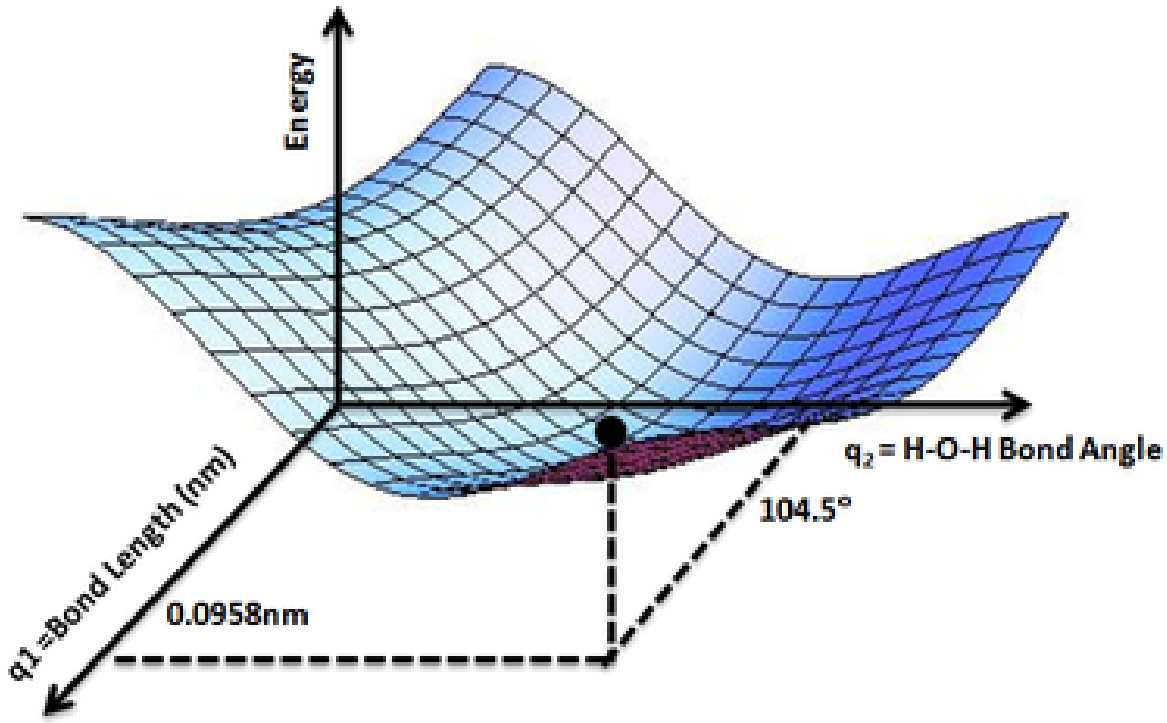

Figure 4: Potential energy surface for ground state of triatomic water molecule with 3 degrees of freedom. Graph show 2 of the 3 degrees of freedom. The global minimum is labeled which gives the stable structure of water. Even for a triatomic molecule, the PES is complex and non-trivial (AimNature 2013)

Machine learning algorithms are proposed for making significantly less computationally expensive calculations. Morawietz et al. (2016) reproduced DFT calculations using ML while employing a previously existing van der Waals correction to calculate the potential energy surface. The result was that the ML neural network was able to reproduce DFT calculation using less computational power and incorporating van der Waals forces. Other techniques for predicting optimized geometry and determining global and local minima on complex PES have used deep ML programs (Cova and Pais 2019). Additionally, ML has been used for prediction of force fields, and thermodynamic, kinetics calculations and the influence of those properties on molecular structure (Cova and Pais 2019).

Rupp et al. (2012) trained a deep ML on atomization energies for a set of compounds in a database. The composition and configuration of the molecules based on coulombic potential and potential energy of each individual atom were used to build a highly intricate algorithm to determine the potential energy for each individual. The algorithm is able to compare the differences between the unknown molecule and the known molecules. The prediction of the atomization energy of unknown molecules by the algorithm involves comparison to known molecules of similar constituent atoms, a weighting function is applied to capture the similarities between the known molecules and the unknown molecule. The algorithm builds a landscape to allow for the prediction of its atomization energy. Impressively, the ML algorithm reproduced the results in milliseconds while DFT takes an average time of one hour. The mean error was $10\frac{kcal}{mol}$ which is within chemical accuracy and has improved accuracy of hybrid DFT (Cova and Pais 2019).

On the other hand molecular dynamic simulation use force field minimization for systems that range from hundreds to thousands of atoms. These traditional approaches require extensive computational resources to obtain reliable data. Cheaper alternatives are desirable and sought after. Cheap molecular dynamics approaches are being developed using ML. To aid in the accuracy and development of the programs, creative techniques are being applied. Molecular dynamic simulations run on a set time interval. Combining traditional methods and ML, allowing some time steps to be used as a training phase for ML to estimate consecutive ones, would allow for more training data. Using such a tool would mean the computationally efficient ML program could aid in simulating a portion of the times steps (Behler 2016). A decision tool for when the ML should be training and when it should be estimating the next step is used. For the abnormal configuration in which the ML is less trained, the calculated molecular dynamics simulation determines the next time steps (Behler 2016). In such simulations, energies and forces are calculated for the molecule for the atomic configuration along with the movement for each time step. The more the data applied to training the AI, the better it would be at handling these abnormal situations and the decision tool would be seldom utilized.

The state-of-the-art *ab initio* methods like DFT treats atoms explicitly, as a result, calculation with a large number of atoms (more than 100) becomes prohibitive. Therefore, the conventional approach is to use potential energies and intramolecular forces which are estimated by atomic potentials or molecular force fields taking a Newtonian approach. Using ML programs can determine such force fields in an adaptive and robust manner. The physics of interacting molecules can be learned by ML algorithms to allow for a faster and more accurate generation of force fields. The result is a ML generated force field can be effectively applied to different type of atoms and molecules (Cova and Pais 2019).

Coarse-grained modelling approximates larger molecules with simplified representation. Instead of modelling each individual atom types, the molecules are reduced into larger grains that make up the molecule. This allows for much larger molecules to be simulated for larger time-scales for molecular dynamics simulations. This allows for longer simulation timescales from picoscale to nanoscale to see more molecular behaviour. Wang et al. (2019) used ML for coarse-graining resulting in a more accurate potential energy surface than traditional methods which often missed relevant features. Additionally, the ML is able to determine when the models did not make physical sense such as overstretched bonds or overlapping atoms by incorporating energy conservation, energy invariance along with translational motion, etc. into the programming. This was then applied to molecular dynamic simulations.

AI algorithms for quantum computational simulations are still in the early stage of development but are demonstrating significant potential. Applications of these simulations become very compelling when considering the combination of these machine learning algorithms. Using ML to determine the coarse-grain models, simplifying the molecule, can be implemented to the molecular dynamics simulations. Then another ML to determine the potential energy surface and thus molecular configuration. Finally, ML can be used to determine the time-steps which it is suitably equipped to predict. In principle, the result could be multiple ML algorithms used for molecular dynamics simulation making computational power no longer the limiting factor. Such an ML program can simulate large molecules on larger time scales than was previously possible with traditional techniques.

1.2 Molecular Property Prediction

The properties of a molecule are dictated by their nuclear and electronic structure. Stable structures are linked to local or global minima along the potential energy surface, therefore structural information can be extracted from the calculated PES. Using the molecular structure, artificial intelligence allows the property prediction of hypothetical molecules without experiment. Compounds could be screened and their favourable properties determined without having to be synthesized. The ability to screen millions of compounds by an ML software which then outputs the most promising candidates for the discovery of novel drugs and materials.

ML has been applied to identifying molecules and materials with optimal properties such as conductivity, aqueous solubility, bioavailability, and toxicity. This presents applications for pharmaceutical, medicinal, and environmental chemistry as it is now possible to search through millions of compounds, discarding toxic candidates, and selecting compounds with wanted properties (Cova and Pais 2019). Additionally, this can be done safely, economically, and sustainably. ML already is achieving better results than what can be done by conventional computational techniques. Shindo and Matsumoto (2019) used a neural network ML algorithm given the atomic coordinates of the molecule to predict, atomization energy, maximal absorption intensity, excitation energy at maximal absorption, HOMO, LUMO, first excitation energy, ionization potential, electron affinity, and polarizability. It was trained on thousands of organic compounds. The results were efficient, accurate and achieved improved results for all properties over DFT quantum-chemical simulations.

Optimal conditions during synthesis for product yields and rates of reactions are able to be predicted (Ahneman et al. 2018). Reaction prediction and retrosynthesis are tasks that can be dealt with simultaneously. Retrosynthesis is the process of segmenting an organic molecule into smaller individual parts which could be used for synthesis. Using available databases of around a million organic reactions, Segler and Waller (2017) achieved a 97% accuracy for reaction prediction and 95% accuracy for retrosynthesis prediction. This work was used for smaller organic compounds and has yet to be implemented to larger scales. Granda et al. (2018) used an organic synthesis robot connected to an ML algorithm. The robot had access to 18 reagents with access to 969 possible reactions that the robot could perform. After completing 10% of the possible reactions, the remaining reactions were able to be predicted with 86% accuracy. Impressive reaction prediction capabilities have been demonstrated although challenges with reactions unique to the data set in which the ML was trained on remain an issue. The predictive quality of the algorithm should not be misunderstood as the AI understanding reaction mechanisms. It simply outputs the most likely products using the data it already has been trained with.

Machine learning and deep learning techniques are being applied to protein structure prediction. This application would allow for the prediction of proteins that can bind to specific targets and proteins which can catalyze different reactions (Torrisi, Pollastri, and Le 2020). Various techniques have been applied, simplifying protein structure into various forms. This involves depicting the protein in its 1D structure arrangement comprising α-helices and β-pleated sheets. Using ML methods, secondary protein structure prediction has reach state-of-the-art results reaching accuracies of 86%. Another method depicts the 2D projection of the tertiary structure where any 2 amino acids are labelled by their spatial distance (Torrisi, Pollastri, and Le 2020). This method can yield 3D protein structure predictions. The latter was used by Senior, Evans, and Jumper (2020) which used a neural network that could build protein fragments and learn such fragments to be used in future calculations. The program, called “AlphaFold”, was trained on 150,000

proteins. Accurate predictions of protein structure were made for 25 out of the 43 proteins attempted. Although the accuracy of the program could significantly improve, the latter competitor made 3 out of 43 correct predictions. This significant improvement gives hope for accurate protein structure prediction in the near future. These protein structures can be with pre-existing docking prediction software to evaluate whether the shape is conducive to binding to a target receptor.

The reaction prediction of molecules and retrosynthesis can be done with a very high accuracy rate although it lacks versatility when dealing with unusual reactions. Protein folding prediction is showing promise with accurate ML programs, the coming years will see some of these promises realized. These various properties prediction programs can be fed various molecules to be screened or evaluated for their effects.

2 Genetic Programming

Computationally cheap and accurate quantum-chemical simulations using ML allow for fast-paced structure determination as well as enhance molecular dynamic simulations. With the advances in molecular property prediction, which is dictated by structure, comes the ability to evaluate millions of compounds for a selected characteristic. Connecting theses application means millions of randomized compounds can be evaluated and selected leading to novel drugs, materials, proteins, and catalyst. However, using randomized compounds is inefficient and lacks the advantages of having a specifically designed molecule. A mechanical engineer can try millions of random configurations of electronics and metal but the best result will still not be as well-designed as a car.

The most auspicious application of AI for chemistry is combining molecular property design, quantum-chemical simulations, and another AI technique, genetic programming. Genetic programming is able to design novel molecules for a given task. Its name comes from its architectural similarities to the process of evolution.

Alan Turing, in 1948 first proposed a machine intelligence or artificial intelligence that would simulate the evolution process to allow a computer to mimic creative tasks (Copeland 2020). Genetic programming can replace design tasks traditionally done by human creativity and thought. Genetic programming(GP) is firmly related to the Darwinian theory of evolution. For chemical applications, the program works as follows. An initial randomized set of compounds is fed into the program. Each compound is simulated to be evaluated for the desired property. The compound is ranked by its score or "fitness." The compounds which score lower for this property are discarded, replicating the evolutionary mechanism of "survival of the fittest." The ones which achieve a high score are chosen to pass along their molecular structure to the next iteration, "pass genes to offspring." The subsequent iteration of "offspring" can be created by three methods. The "parent" can be cloned, mutated, or combined with another surviving design. This step is defined by the architecture of the GP and implements the "genetic variation" of the program. These variations in structure will affect the evaluated property. The compounds are then evaluated, and the cycle continues for thousands of iterations. The result is a computer program that can create practical, creative, and novel designs.

GP is currently being used to create human-competitive designs for many fields such as optics and electronics. It has created patentable inventions in these fields (Duettmann and Lewis 2018). The appeal of GP is that little human input is needed for the amount of output intelligence. GP is best applied to problems which

humans have little understanding of. Since the human ability to design molecules is limited, we need a program which does not rely solely on the intelligence of human and can, therefore, surpass the best of human creativity. Additionally GP is easily applicable to different fields as inputs architecture can remain in the native format of the field. For chemical applications, this means our 3D coordinate system of modelling molecule is sufficient to feed into the GP.

The effectiveness of GP has strongly correlated with available computer power, meaning, similar to other computational programs, the hardware is an expected limitation. However, the architecture of GP allows for optimization methods. For example, work can be split between separate computers leveraging parallel programming techniques. GP, like evolution, does not find the best possible solution but candidates that are well equipped to excel at the desired property. There are multiple solutions to a problem and so different solutions can be outputted by the program.

3 AI for Molecular Design

Advancements in technology have led to impressive feats of engineering such as skyscrapers and aircraft containing thousands of individual parts. Biology shows us exceptional capabilities that are possible when using intricate molecular-scale designs. Synthetic molecular designs pale in complexity to evolutionary designed molecular systems and larger macro-scale man-made inventions. Humans lack the natural capacity to understand the quantum process due to its counter-intuitive nature, thus we lack the ability to effectively design molecular structures containing thousands of atoms. An alternative method of applying macroscale design ideas to the nanoscale level is bound to be inefficient and highly limiting due to the difference in the physical spaces in which they are implemented. The conclusion is that machine aided design is necessary to overcome the limitations of human creativity towards creating synthetic molecular designs.

The evolutionary process has been proven to achieve complexity and powerful bio-nanochemistry such as DNA replication and protein production using ribosomes. For this reason, using a machine learning genetic program, which mimics evolution, is not biased by macroscale realities and presents exciting applications to advance to the ability of synthetic design. The connection of ML for quantum-chemical simulations and molecular property prediction programs would allow for the creation of such a AI for molecular design (Figure 5). What is proposed is a general, efficient, modular AI framework that produces molecular designs for nanotechnology research. Connecting these individual AI platforms into one encompassing software would lead to novel and unexpected discoveries with far-reaching applications.

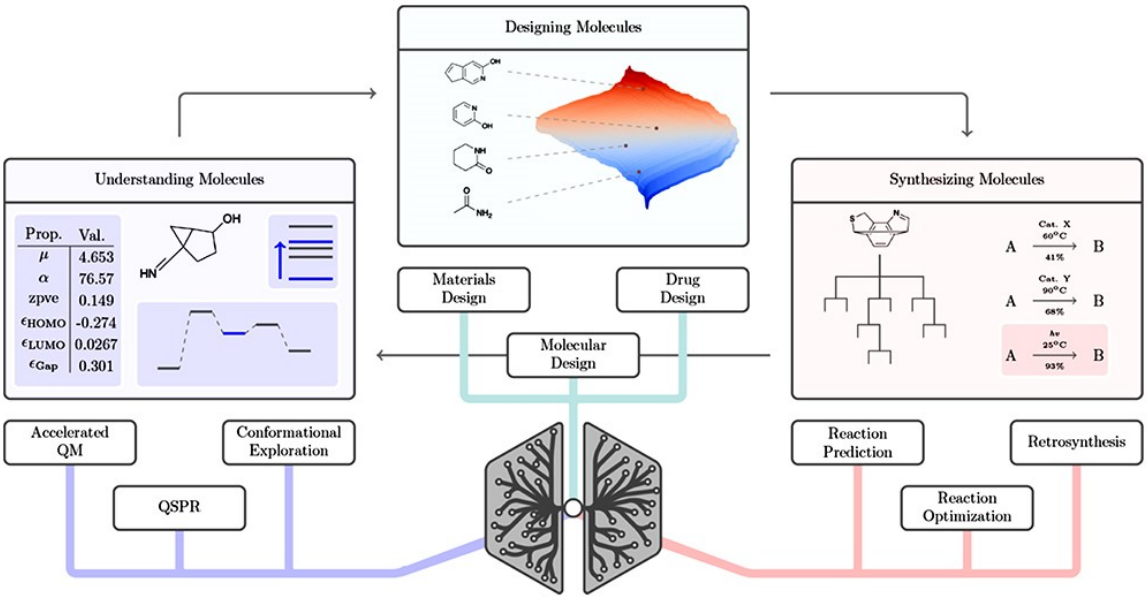

Figure 5: Contribution of Deep ML algorithms for chemical problems and the connection which can be used for additional applications (Cova and Pais 2019)

As proposed by Duettmann and Lewis (2018), the architecture of such a tool would consist of seven parts: simulator, constraints, specifications, GP, visualizer, evaluator, and control interface. Both the simulator and evaluator would implement ML programs discussed in section 1. A quantum-chemical simulator is necessary for ensuring physically viable molecules and structures are being evaluated. As discussed, using traditional methods, these simulators present computational limitations, but this challenge could be overcome by using ML techniques. As discussed in section 1.1, ML for emulating DFT calculations would allow for cheap computational cost for needed simulations of static molecules. Further, the implementation of molecular dynamic simulations would allow for the design of moving molecules and reaction-based designs.

Constraints for the AI such as the physical space which the simulations are dealing with would be able to be implemented. Examples include, constraints regarding the environment which the AI is investigating designs for or constraints for the makeup of the design itself, such as using only carbon atoms or only amino acids. Specifications would allow for the designs to meet certain requirements, as well as define the desired properties which the GP will search for. The specifications are then loaded into the evaluator which is the function that will score each of the candidates for each iteration. The evaluator will assign a better score to the molecules that better demonstrate the desired property.

The evaluator would implement molecular property prediction programs discussed in section 1.2 to be able to determine which molecules are optimal for chosen properties. This means properties for electronic data such as absorption, polarizability and conductivity could be evaluated, and thus designed for. Additionally, for pharmaceutical applications, bio-availability and toxicity could be evaluated. Reactive properties and catalyst for optimizing production yields and reaction rates could as well be searched. Lastly, molecules and amino acid protein backbones could be evaluated for structure and docking ability with receptors. The AI GP algorithm will then iteratively search for designs that satisfy the specifications and attempt to maximize the evaluator score as explained in section 2. A visualizer to be able to see the progress of the AI, and a

control interface to access modular components would also be implemented to the software.

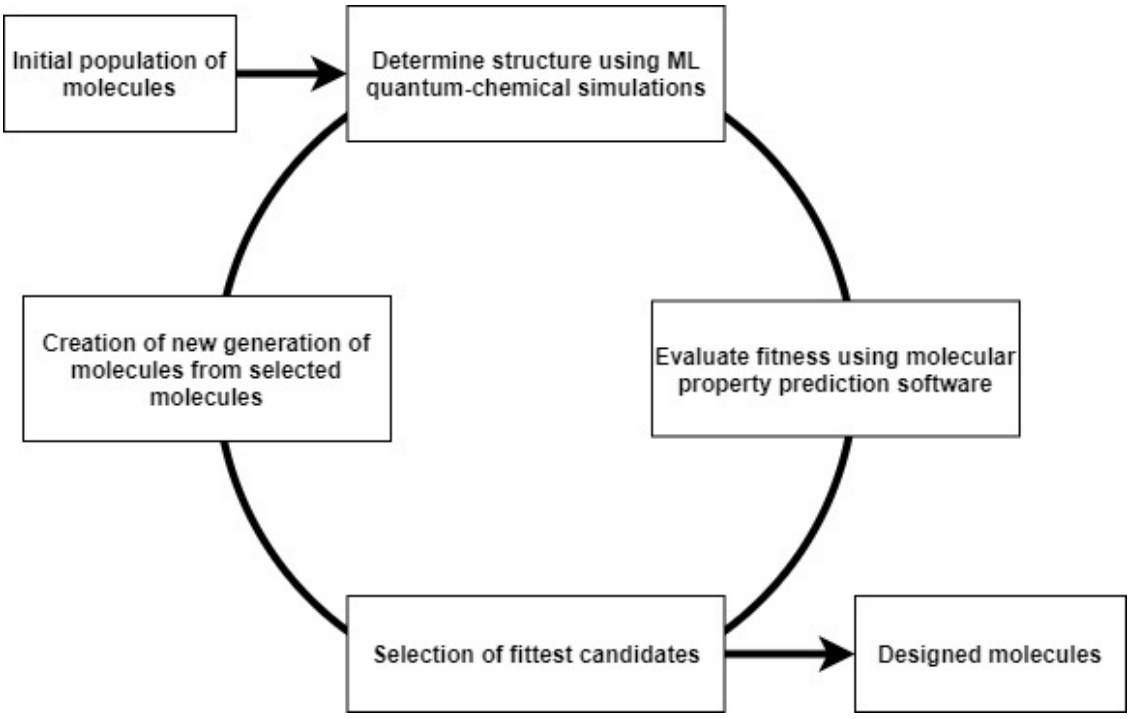
Initial population of molecules
Determine structure using ML quantum-chemical simulations
Evaluate fitness using molecular property prediction software
Selection of fittest candidates
Designed molecules
Creation of new generation of molecules from selected molecules

Figure 6: Diagram of steps for one generation done by AI for molecular design

Additional components can be added to the AI. Since quantum-mechanical simulations are computationally expensive, a cheaper tool to estimate the quality of the design could be used prior to simulation to ensure that running the simulations is worthwhile (Duettmann and Lewis 2018). Example include using a cheaper tool to first confirm that designs are physically viable or confirming that design are novel to the AI and therefore do not need to be re-simulated. The usefulness of such a program would be enhanced by incorporating an ML that can determine and evaluate the synthesis procedure for the candidate molecules. Using ML techniques discussed in section 1.2, the algorithm could determine a method of synthesis for the design. This procedure could then be evaluated along with the molecular properties. This would ensure the final candidates are synthesizable, and such a procedure would be determined by the AI.

Artificial intelligence programs for a molecular design using GP are currently being developed. MoleGear, a Java-based program attempted to design inhibitor molecules for the immunodeficiency virus 1 (HIV-1) protease (Chu and He 2019). The GP used a small population of 100 and only 30 generations and a docking software to evaluate candidates. The designed molecule contained many fragments and structural similarities to indinavir, a proven treatment for HIV, as seen in Figure 7. The conclusion is that MoleGear can produce novel chemically feasible drug-like designs using genetic programming and a docking software (Chu and He 2019).

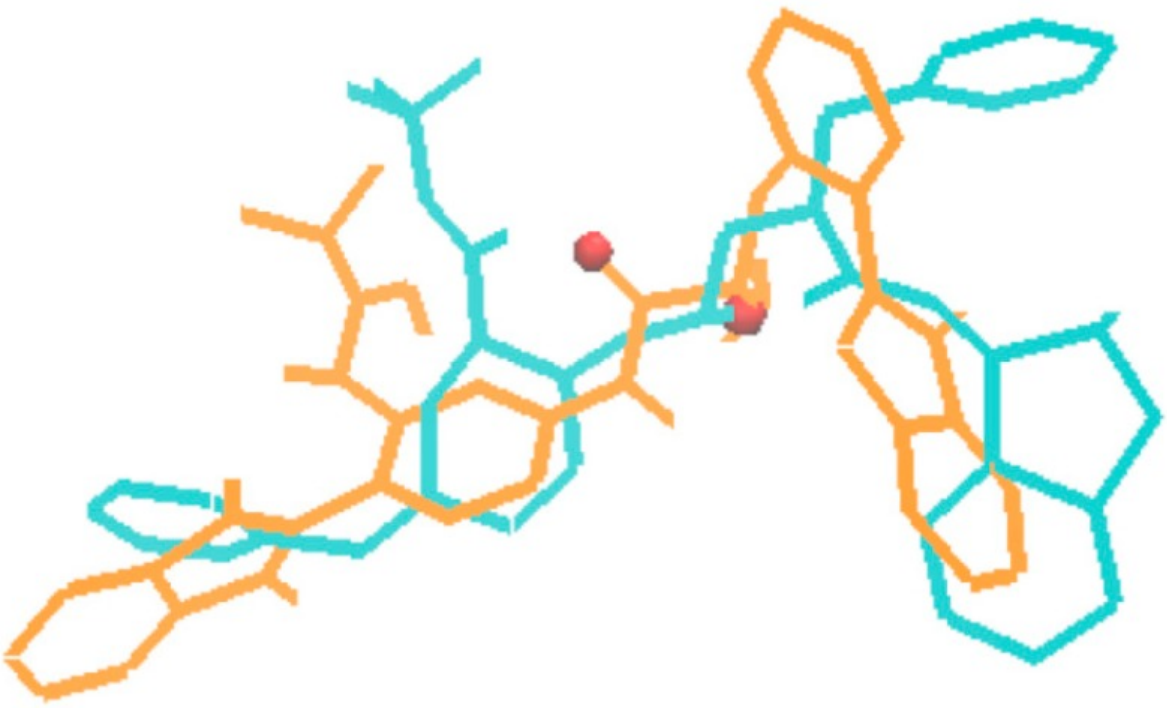

Figure 7: Superposition of Molegear created design (orange) using genetic programming and indinavir (cyan) (Chu and He 2019)

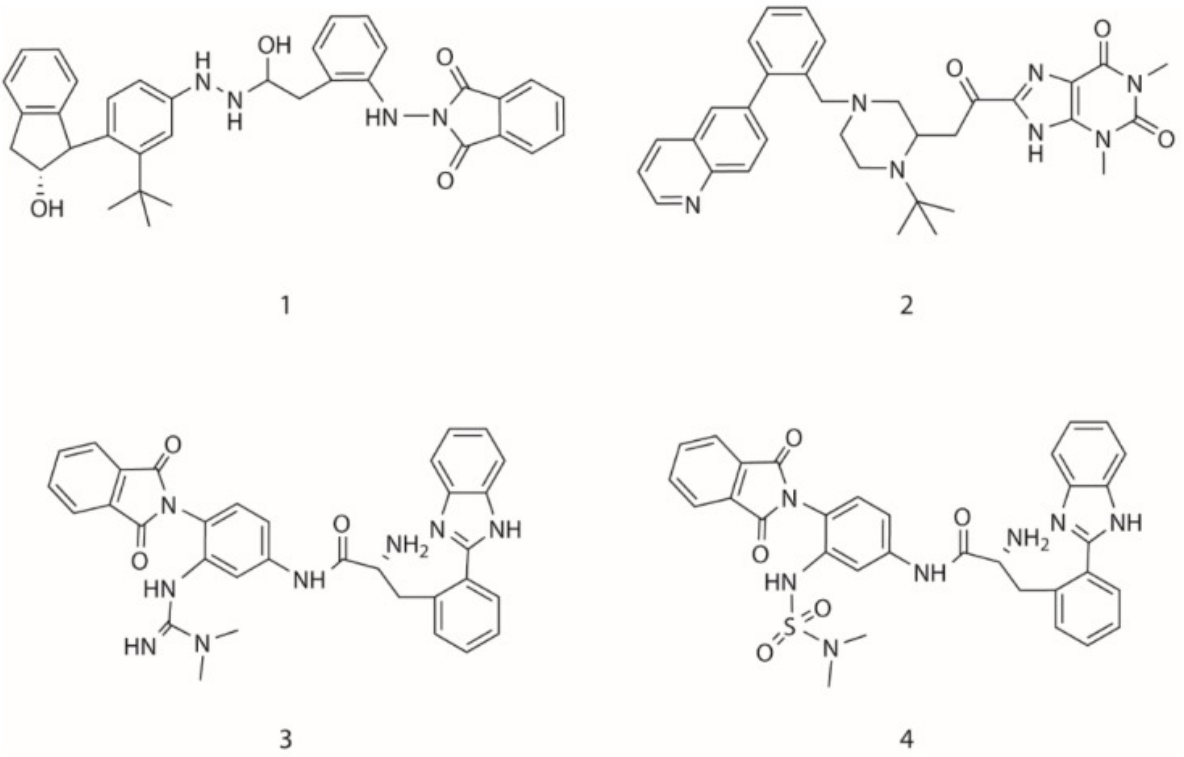

Figure 8: Candidate HIV-1 inhibitor molecules designed by Molegear by genetic programming (Chu and He 2019)

The development of new catalysts by genetic programming has also been attempted. Genetic programming has been used to design new catalysts using computationally cheap ML algorithms to screen for compounds. This allows for the target design of novel catalyst which exhibit certain properties (Freeze, Kelly, and Batista 2019). These methods are sure to be expanded for more efficient and robust AI for molecular design. Using this AI method, the overall cost and time are reduced for discovery and development. By using an AI which can create targeted designs, it is expected superior catalyst will be produced than ones designed by the traditional largely trial and error approach. However, these AI programs are still in there infancy and require further improvement for future work described in section 5.

4 AI for Molecular Design Applications

As described, researchers are developing ML optimised to solve specific chemical problems very well. The unification of many of these programs present significant applications to an AI for molecular design. As concurrent research progresses in these individual fields within property prediction, quantum simulation, and genetic programming, the combination of these technologies into an Artificial Intelligence for Molecular Design would yield near-endless possibilities and applications. Applying evolutionary tactics would allow for a molecular design process proven by our observations of biology. Perhaps more exciting, this AI would be a tool for molecular design which is not limited by human chemical knowledge and research, thus resulting in a completely new generation of complex and effective molecular designs. All the parts of the AI are identified and have already been used for the design of successful drug candidates and catalysts.

Relatively simple examples of tasks that the AI could accomplish would be searching for a molecule that absorbs a certain wavelength of light, a molecule that has a high polarity, or a molecule which is able to optimally filter out other molecules according to size. Adding more complexity would be searching for a molecule with the desired structure which will bind with a receptor while being non-toxic, or non-reactive. Implementation of ML techniques for protein structure and docking would enable the rapid development of better pharmaceuticals. Once a receptor is determined as a target for a compound to bind to, the structure of the receptor location can be modelled and the AI can search for molecules with a structure that will bind to such receptor by using docking prediction software. The output would be applicable to structural candidates that have also been screened on their toxicity, bioactivity, and synthesizability. Currently, researchers spend years or even decades determining appropriate structures to bind to such a receptor through largely inefficient processes, often attempting to tweak changes in molecules. Such research is currently being applied to search for many treatments such as for Alzheimer's (Mehrazma and Rauk 2019). This task and subsequent testing and synthesis development tasks could be done cheaply and efficiently by an AI molecular design software. Since GP has the ability to test and develop intelligently designed molecules, it is very feasible that candidates would far exceed the functionality of parallel human-designed compounds. Time spent on synthesizing and screening would be greatly reduced as far fewer compounds would go through and fail experimental testing and health screenings.

Enzymes, catalysts built by proteins, can have trillion-fold effects on reaction rate (Duettmann and Lewis 2018). Such technology is far out of reach of human capabilities due to our limited ability for molecular design. With the implementation of molecular dynamic simulation, it is conceivable that novel enzymes for new reactions could be designed by such software. Needed would be a designed fitness function related to the rate at which the reactants are formed into the products and specifications to only amino acid fragments

for the designed molecule. The output would be a synthetic enzyme that could be deconstructed into its core amino acid chain. This could then be used as the script for a designed RNA string. Then using the designed RNA, the enzyme could be produced by the ribosomes in our cells.

The future of nanomachine research will largely rely on such an AI for molecular design. The current state-of-the-art designs are embryonic in comparison to nanomachines found in the body. The limitations of human creativity to understand the quantum systems are further amplified when considering complex dynamic molecules. The molecular machines found in biology demonstrate techniques and designs far out-reaching current complexity and capabilities demonstrated by today's synthetic attempts. Cellular organelles demonstrate the physical viability that advanced nanotechnology could one day be achieved. Ribosomes are kin to molecular assembly lines that can build molecules with precision, speed, and reliability. The ribosome contains 40-80 molecules and around 250,000 atoms (Britannica 2016). This molecular structure, "designed" by the process of evolution demonstrates the far-reaching potential of genetic programming for molecular design. With a fully developed artificial intelligence for molecular design, by inputting relatively simplistic specification such as the presence of mRNA and transfer RNA, and developing a fitness function for reading and transcription, an effective multimolecular, dynamic structure could be outputted which would complete the same task as a ribosome. The understanding of how this structure functions and completes its task may be completely unknown, yet is still able to be designed. With an advance AI for molecular design, complex molecular structures could be designed which rival the far-out-of-reach capabilities of biochemistry.

An advanced AI for molecular design could be able to iteratively search for molecules with high complexity of movement to give rise to a new generation of molecular design with surely unpredictable applications. The AI automatically finds nanomachine candidates that would use the molecular movement to aid in the task being searched for or could be manually encouraged to develop moving molecules. The transition from human hypothesized chemistry to AI developed techniques would possibly be one of the most significant changes to the field to date.

5 Future Work

The current success of AI for molecular design has brought programs that can apply genetic programming to create novel designs. Chemically viable pharmaceutical candidates have been designed using GP (Chu and He 2019). GP has also been used for the targeted design of novel catalysts (Freeze, Kelly, and Batista 2019). However, these programs are in their beginning stages, therefor presenting plenty of future research opportunities.

An AI for molecular designs would incorporate three separate ML components. The simulator requires ML based quantum-chemical simulations, the evaluator requires ML based molecular property prediction programs, and the ML GP itself. My future contribution is to work in these 3 areas.

Most current training data comes from theoretical simulations. The implementation of already available experimental data for training of the ML would provide a reliable and accurate addition. Increased reliable data would have a significant effect on the effectiveness of the program. Cases will exist where experimental data is not available, and DFT will not deliver reliable results. DFT is a single reference method and therefore breakdown for cases that have degeneracy. As well, DFT does not take into account non-bonding

interactions such as van der Waals forces etc. The implementation of *ab initio* methods such as self-consistent field (SCF), multi-reference configuration interaction (MRCI) etc. could be used to ensure the accuracy of the AI base on the training data set with-respect-to these abnormal cases. This would make training more computationally expensive however the AI would benefit in the end from better accuracy.

Researchers have trained their various ML programs to excel at specific chemical problems. A significant amount of the work which has to be done is the connection of these various ML programs into an AI for molecular design. This involves selecting ML programs for the simulator and evaluator. Current work has not yet thoroughly benchmarked ML based quantum-chemical simulations nor ML based property prediction programs. For both, ML techniques significantly reduce the computational cost, and in some cases, improve accuracy over traditional approaches. GP in other fields has shown to show a strong correlation between results and computing power (Duettmann and Lewis 2018). More computing power allows for a larger population and more generation to be simulated within the GP. Chu and He (2019) was able to produce a competent design using a tiny population of 100 and just 30 generations. Million fold optimization was demonstrated when using ML for quantum-chemical simulations over traditional techniques. By applying ML simulation techniques, momentously larger population and generation sizes could be produced and in turn large advances in the effectiveness of the AI.

References

[Ahn+18] Derek T. Ahneman et al. "Predicting reaction performance in C–N cross-coupling using machine learning". In: 360 (6385 2018), pp. 186–190. DOI: 10.1126/science.aar5169.

[Aim13] AimNature. *Potential Energy Surface for Water*. 2013. URL: https://commons.wikimedia.org/wiki/File:Potential_Energy_Surface_for_Water.png.

[alp19] alphabold. *Neural Networks anf Depp Learning: An Overview*. Dec. 2019. URL: https://alphabold.com/neural-networks-and-deep-learning-an-overview/.

[Beh16] Jörg Behler. "Perspective: Machine learning potentials for atomistic simulations". In: *The Journal of Chemical Physics* 145.17 (2016), p. 170901. DOI: 10.1063/1.4966192. eprint: https://doi.org/10.1063/1.4966192. URL: https://doi.org/10.1063/1.4966192.

[Bri16] Encyclopaedia Britannica. "Ribosome". In: (Sept. 2016). URL: https://www.britannica.com/science/ribosome.

[CH19] Yunhan Chu and Xuezhong He. "MoleGear: A Java-Based Platform for Evolutionary De Novo Molecular Design". en. In: *Molecules* 24.7 (Jan. 2019), p. 1444. DOI: 10.3390/molecules24071444. URL: https://www.mdpi.com/1420-3049/24/7/1444 (visited on 08/17/2020).

[Cop20] B.J. Copeland. "Alan Turing And The Beginning Of AI". In: (Aug. 2020). URL: https://www.britannica.com/technology/artificial-intelligence.

[CP19] Tânia F. G. G. Cova and Alberto A. C. C. Pais. "Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns". In: *Frontiers in Chemistry* 7 (2019), p. 809. ISSN: 2296-2646. DOI: 10.3389/fchem.2019.00809. URL: https://www.frontiersin.org/article/10.3389/fchem.2019.00809.

[DL18] Allison Duettmann and James B. Lewis. *Artificial Intelligence for Nanoscale Design*. 2018. URL: https://foresight.org/wp-content/uploads/2018/11/AI-for-Nanoscale-Design-report-2.pdf.

[FKB19] Jessica G. Freeze, H. Ray Kelly, and Victor S. Batista. "Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists". In: *Chemical Reviews* 119.11 (2019). PMID: 31059236, pp. 6595–6612. DOI: 10.1021/acs.chemrev.8b00759. eprint: https://doi.org/10.1021/acs.chemrev.8b00759. URL: https://doi.org/10.1021/acs.chemrev.8b00759.

[Gra+18] Jarosław M. Granda et al. "Controlling an organic synthesis robot with machine learning to search for new reactivity". en. In: *Nature* 559.7714 (July 2018), pp. 377–381. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0307-8. URL: http://www.nature.com/articles/s41586-018-0307-8 (visited on 08/17/2020).

[Hau16] Kyle Hauptfleisch. *This Is How Artificial Intelligence Will Shape eLearning For Good*. 2016. URL: https://elearningindustry.com/artificial-intelligence-will-shape-elearning.

[Mor+16] Tobias Morawietz et al. "How van der Waals interactions determine the unique properties of water". In: *Proceedings of the National Academy of Sciences* 113.30 (2016), pp. 8368–8373. ISSN: 0027-8424. DOI: 10.1073/pnas.1602375113. eprint: https://www.pnas.org/content/113/30/8368.full.pdf. URL: https://www.pnas.org/content/113/30/8368.

[MR19] Banafsheh Mehrazma and Arvi Rauk. "Exploring Amyloid-β Dimer Structure Using Molecular Dynamics Simulations". In: *The Journal of Physical Chemistry A* 123.22 (2019). PMID: 31082235, pp. 4658–4670. DOI: 10.1021/acs.jpca.8b11251. eprint: https://doi.org/10.1021/acs.jpca.8b11251. URL: https://doi.org/10.1021/acs.jpca.8b11251.

[Rup+12] Matthias Rupp et al. "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". In: *Phys. Rev. Lett.* 108 (5 Jan. 2012), p. 058301. DOI: 10.1103/PhysRevLett.108.058301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.108.058301.

[SEJ20] A.W. Senior, R. Evans, and J. Jumper. "Improved protein structure prediction using potentials from deep learning". In: *nature* 577 (2020), pp. 706–710. URL: https://doi.org/10.1038/s41586-019-1923-7.

[SM19] Hiroyuki Shindo and Yuji Matsumoto. *Gated Graph Recursive Neural Networks for Molecular Property Prediction*. 2019. arXiv: 1909.00259 [cs.LG].

[SW17] Marwin H. S. Segler and Mark P. Waller. "Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction". In: *Chemistry – A European Journal* 23.25 (2017), pp. 5966–5971. DOI: 10.1002/chem.201605499. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/chem.201605499. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/chem.201605499.

[TPL20] Mirko Torrisi, Gianluca Pollastri, and Quan Le. "Deep learning methods in protein structure prediction". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1301–1310. ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2019.12.011. URL: http://www.sciencedirect.com/science/article/pii/S2001037019304441.

[Wan+19] Jiang Wang et al. "Machine Learning of Coarse-Grained Molecular Dynamics Force Fields". In: *ACS Central Science* 5.5 (2019), pp. 755–767. DOI: 10.1021/acscentsci.8b00913. eprint: https://doi.org/10.1021/acscentsci.8b00913. URL: https://doi.org/10.1021/acscentsci.8b00913.